2Н

УДК 004.8

Н.А. Новоселова[1], А.С.Мастыкин[2], И.Э. Том[1]
[1]Объединенный институт проблем информатики НАН Беларуси, г. Минск, Беларусь
[2]Белорусский государственный медицинский университет, г. Минск, Беларусь
novosel@newman.bas-net.by

Эволюционный подход к выделению информативных признаков в задачах анализа медицинских данных

В статье рассматривается подход к выделению информативных признаков применительно к задаче распознавания подтипов транзиторных ишемических атак. Согласно предложенному подходу задача выделения признаков рассматривается как задача многокритериальной оптимизации с двумя критериями. Оптимизация осуществляется с использованием специального генетического алгоритма, позволяющего в процессе эволюции получить множество недоминируемых решений оптимизационной задачи. Предложенный подход позволяет подключить эксперта на этапе окончательного принятия решений, предоставляя ему возможность отбора подмножества признаков, наиболее соответствующего его знаниям и представлениям о решаемой задаче.

Введение

Одним из важнейших этапов процесса извлечения знаний из большого объема накопленных медицинских данных является этап предобработки исходных данных, включающий выделение информативных признаков. Благодаря широкому распространению компьютерных технологий, в базах данных медицинских учреждений накапливается большое количество разнородной медицинской информации, большая часть которой напрямую не связана с решением какой-либо конкретной задачи, как например задачи классификации или прогноза. В этом случае исключение из рассмотрения избыточных и несущественных признаков позволяет не только повысить точность решения задачи и сократить время на поиск решения, но и получить более простой и понятный результат [1].

В данной статье рассматривается применение эволюционного подхода к выделению признаков для дифференциальной диагностики подтипов транзиторных ишемических атак (ТИА). Исходными данными в этом случае является набор клинических и персональных признаков, характеризующих пациента, который в свою очередь относится к одному из четырех классов [2]. Параллельно с процессом выделения признаков решается задача классификации, целью которой является предсказание класса для конкретного объекта данных, основываясь на значениях предсказывающих признаков. Задача выделения признаков рассматривается как оптимизационная задача с двумя оптимизируемыми критериями: минимизация ошибки классификации и количества отобранных предсказывающих признаков. Для решения этой задачи предлагается использовать специально разработанный генетический алгоритм (ГА) для многокритериальной оптимизации [3]. Основное преимущество применения этого алгоритма для выделения признаков состоит в

возможности получения множества оптимальных решений с учетом двух критериев, так называемых недоминируемых решений многокритериальной оптимизационной задачи. Такой подход позволяет избежать изначального жесткого определения весовых коэффициентов для отдельных критериев. Выбор окончательного множества предсказывающих признаков из различных недоминируемых комбинаций может осуществляться либо экспертом согласно его знаниям и опыту, либо автоматически с использованием тестового набора данных.

1. Задача многокритериальной оптимизации

Большинство решаемых практических задач предполагают поиск решения, являющегося оптимальным согласно нескольким критериям. Однако большинство методов, используемых для решения этих задач, использует единый, составной оптимизируемый критерий. В этом случае задача многокритериальной оптимизации сводится к одной или нескольким задачам однокритериальной оптимизации. Существует огромная разница между двумя этими задачами. При однокритериальной оптимизации осуществляется поиск единственного оптимального решения. При многокритериальной оптимизации осуществляется поиск нескольких оптимальных решений, что позволяет равным образом учитывать все оптимизируемые критерии [3]. После завершения оптимизации пользователь имеет возможность выбрать наилучшее с его точки зрения решение, представляющее собой компромисс между несколькими противоречивыми критериями.

Поиск множества решений при многокритериальной оптимизации основывается на концепции Парето-оптимальности. Основная ее идея заключается в определении понятия недоминируемости для отдельных решений оптимизационной задачи. Решение x_1 доминирует другое решение x_2, если одновременно выполняются два следующих условия:

1. Решение x_1 не хуже решения x_2 по любому из рассматриваемых в задаче критериев.
2. Решение x_1 строго лучше решения x_2 по крайней мере по одному из критериев.

Если не существует ни одного решения, удовлетворяющего вышеперечисленным условиям, то x_2 является недоминируемым или Парето-оптимальным решением многокритериальной задачи.

Согласно предложенному в статье подходу, выделение информативных признаков для решения задачи распознавания подтипов ТИА представляется как задача многокритериальной оптимизации.

Пусть Ω – множество различных подмножеств признаков, характеризующих объект данных. Каждое подмножество представляет собой некоторую комбинацию входных признаков из максимально возможного количества комбинаций 2^n, где n – количество входных признаков. Требуется выделить подмножество признаков $S \in \Omega$, которое является решением двухкритериальной задачи оптимизации с двумя следующими критериям:

$$\max f_1(S), \min f_2(S), \quad (1)$$

где $f_1(S)$ – количество правильно классифицированных объектов с использованием подмножества признаков S, $f_2(S)$ – количество элементов подмножества признаков S.

В предлагаемом подходе для выделения признаков использован генетический алгоритм, который имеет модифицированную схему реализации применительно к

задаче многокритериальной оптимизации. Согласно алгоритму не требуется изначальное определение весовых коэффициентов, соответствующих отдельным целевым критериям [3]. Решение задачи оптимизации в этом случае можно получить в виде нескольких недоминируемых подмножеств признаков.

Для расчета точности классификации с использованием подмножества признаков используется алгоритм k-ближайших соседей [4]. Согласно этому алгоритму для каждого объекта определяется k-ближайших соседей в пространстве признаков. Выбор соседей обычно выполняется на основании значений евклидовых расстояний, хотя можно использовать другие метрики (например, расстояние Махаланобиса). В качестве класса объекта выбирается класс, к которому относится большинство из k-ближайших соседей.

2. Описание эволюционного подхода

Среди различных категорий алгоритмов выделения признаков генетические алгоритмы стали применяться относительно недавно. Генетические алгоритмы представляют собой стохастические методы решения оптимизационных задач, в основе которых лежит моделирование процессов биологической эволюции [5]. Генетические алгоритмы можно отнести к наиболее эффективному методу глобального поиска в многомерном пространстве признаков, позволяющему получить оптимальное или близкое к нему решение поставленной задачи и учесть возможные взаимозависимости между признаками. Многие авторы применяли генетические алгоритмы для отбора признаков, где в качестве значения оценочной или оптимизируемой функции выступала точность классификации с использованием дерева решений и классификаторов, основанных на принципе ближайших соседей [6], [7].

В работе [6] описан один из первых подходов к использованию генетического алгоритма для отбора признаков. В [6] ГА используется для поиска оптимального бинарного вектора, где каждый бит соответствует отдельному признаку (рис. 1). Если i-й бит вектора равен единице, то соответствующий ему признак участвует в классификации; если бит равен нулю, тогда соответствующий признак исключается из дальнейшего анализа.

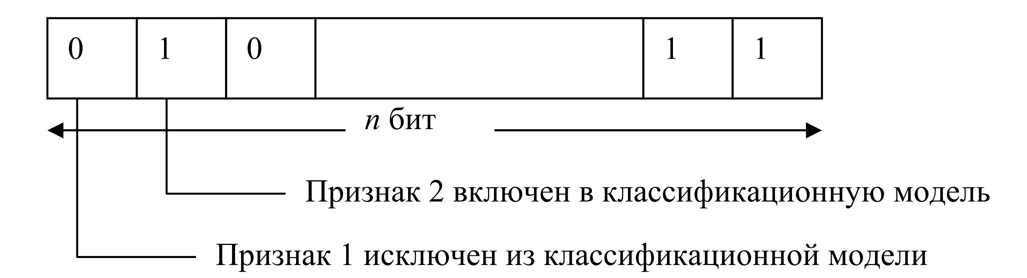


Рисунок 1 – n-мерный бинарный вектор, определяющий особь популяции генетического алгоритма для отбора признаков

Основными предварительными этапами при использовании генетического алгоритма для выделения информативных признаков является определение кодировки особей, оценочной функции или функции приспособленности и основных операций селекции и рекомбинации.

В общем случае ГА может осуществлять поиск наилучшей диагональной матрицы **W** или вектора ее диагональных элементов $\overline{w} = [w_1, w_2, \dots, w_n]$, который представляет собой «наилучшее» преобразование исходного признакового пространства с целью максимизации/минимизации оптимизируемого критерия. Для расчета значения оптимизируемого критерия каждый входной объект данных $\overline{x} = [x_1, x_2, \dots, x_n]$ преобразуется с использованием генетически сгенерированной матрицы **W** с целью получения нового вектора признаков $\overline{y}$:

$$\overline{y} = W(\overline{x}),$$

где $W = \begin{bmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & w_n \end{bmatrix}$,

a) если $w_i \in \{0,1\}, 1 \le i \le n$, то в качестве элементов вектора $\overline{w}$ используются только бинарные значения. В этом случае если i-ый компонент вектора $\overline{w}$ равен единице, то i-ый признак сохраняется в отбираемом подмножестве, в противном случае признак исключается из подмножества. В этом случае осуществляется отбор признаков и сокращается размерность исходного признакового пространства;

b) если $w_i \in [a,b]$, например $w_i \in [0,10], 1 \le i \le n$, то осуществляется выделение признаков, т.е. происходит поиск относительных весов признаков, которые обеспечивают наилучшее значение оптимизируемого критерия. Значения весов признаков определяют их полезность для решения соответствующей оптимизационной задачи. Весовые коэффициенты со значениями, близкими к нулю, указывают на низкую информативность признака. В этом случае эти признаки могут быть исключены из рассмотрения;

c) если в особи ГА закодировать как вектор с бинарными значениями, так и вектор весовых коэффициентов, то возможно одновременно решить задачу линейного масштабирования (взвешивания) и отбора признаков, что позволяет определить не только состав отобранных информативных признаков, но и степень их информативности для конкретной задачи.

Таким образом, предложенный в настоящей работе эволюционный подход включает два способа выделения информативных признаков:

1) отбор некоторого количества предсказывающих классификационных признаков из всего множество анализируемых признаков;

2) взвешивание признаков с одновременным отбором.

Для каждого из перечисленных способов используется различное кодирование особей генетического алгоритма и соответственно различные операции рекомбинации и мутации.

В связи с тем, что выделение признаков рассматривается как задача многокритериальной оптимизации, то приспособленность каждой особи генетического алгоритма определяется двумя численными значениями: точностью классификации набора данных с использованием алгоритма k-ближайших соседей и количеством выделенных признаков. Основные генетические операции используемого ГА описаны в работе [3].

Генетический алгоритм для многокритериальной оптимизации позволяет на каждой генерации выделять все недоминируемые решения и передавать их в следующую генерацию, тем самым, обеспечивая сохранение наиболее приспособленных особей в последующих поколениях и сходимость ГА.

3. Описание исследуемого набора данных и результатов вычислений

Исходный исследуемый набор данных состоит из 101 наблюдения клинически выверенных случаев пациентов с атеротромботическим этиопатогенезом эпизодов ТИА (СубТИА1) – 22 наблюдения, кардиоэмболическим (СубТИА2) – 23 наблюдения и гипертензивным (СубТИА3) – 22 наблюдения. Контрольная группа НОРМА включала 34 наблюдения. Каждое наблюдение характеризуется измерениями по 25 клиническим и персональным признакам.

3.1. Результаты работы ГА в случае отбора признаков

Рассмотрим результаты работы ГА в случае решения задачи отбора признаков для распознавания подтипов ТИА. Используемые в этом случае значения параметров ГА приведены в табл. 1. Каждая особь ГА представляет собой частное решение задачи отбора признаков и состоит из n ген, где n – количество всех рассматриваемых признаков ($n = 25$). Каждый ген может принимать значение 0 или 1, что указывает на исключение/включение соответствующего признака в состав подмножества отбираемых признаков.

Таблица 1 – Значения параметров ГА

Параметр	Значение
Размерность популяции ГА	200
Количество генераций	100
Вероятность рекомбинации	0,8
Вероятность мутации	0,1

На рис. 1 представлены значения ошибки классификации, с использованием подмножеств признаков различной размерности, полученных в ходе работы ГА.

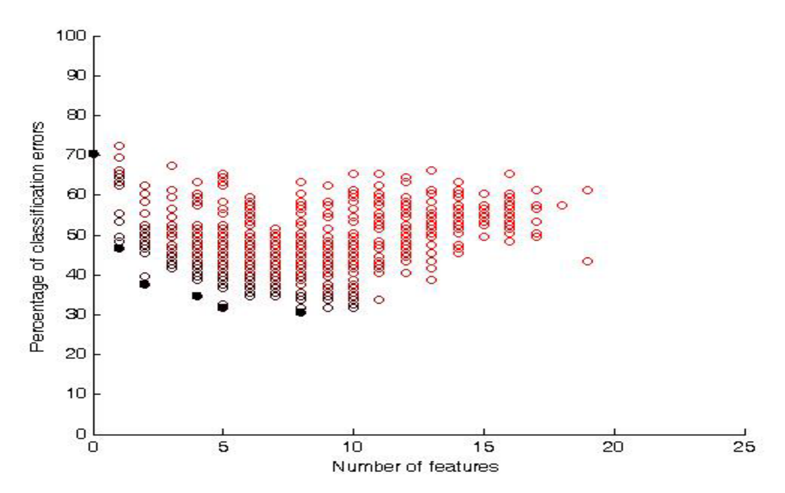


Рисунок 1 – Эволюция популяций ГА в двухмерном пространстве оптимизационных критериев

Полученные в процессе работы ГА недоминируемые решения – Парето-оптимальный фронт – многокритериальной оптимизационной задачи отбора признаков представлены на рис. 2.

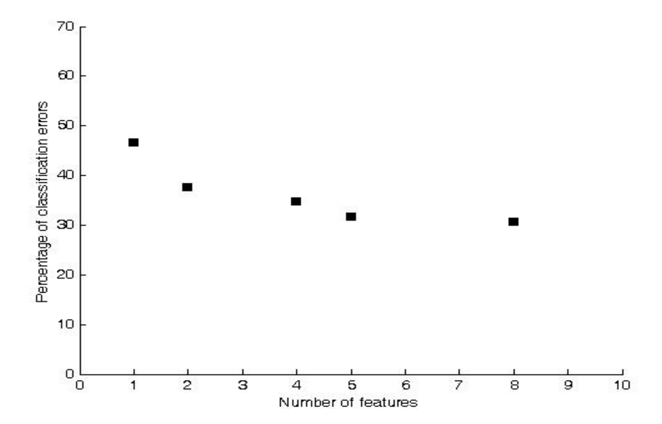


Рисунок 2 – Недоминируемые подмножества признаков: горизонтальная ось определяет количество признаков, вертикальная – ошибку классификации (распознавания) подтипов ТИА

Согласно рис. 2 подмножество признаков, обеспечивающее минимальную ошибку классификации (≈ 30 %) с использованием алгоритма k-ближайших соседей состоит из 8 признаков.

3.2. Результаты работы ГА в случае взвешивания признаков с одновременным отбором

Рассмотрим результаты работы ГА в случае решения задачи взвешивания признаков с одновременным отбором для распознавания подтипов ТИА. Используемые в этом случае параметры ГА идентичны приведенным в табл. 1. Каждая особь ГА представляет собой частное решение задачи взвешивания и отбора признаков и состоит из 2*n ген, где n – количество всех рассматриваемых признаков ($n = 25$). Первые n ген могут принимать действительное значение w_i в интервале [0,10], обеспечивающее независимое линейное масштабирование отдельных признаков, последующие n ген могут принимать бинарные значения и предназначены для отбора признаков. Таким образом, результатом работы ГА являются недоминируемые подмножества признаков с весовыми коэффициентами (рис. 3).

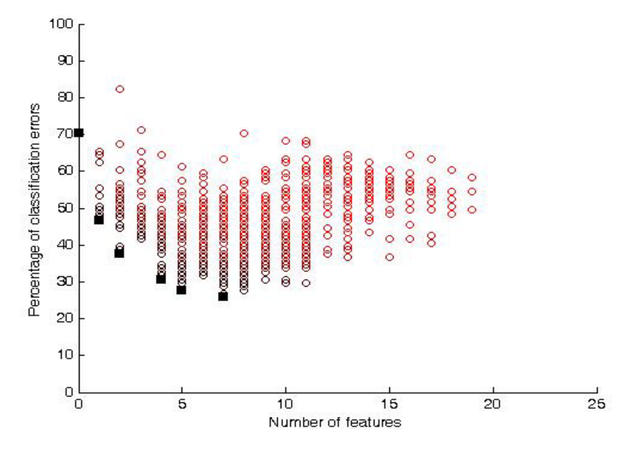


Рисунок 3 – Эволюция популяций ГА в двухмерном пространстве оптимизационных критериев

Полученные в процессе работы ГА недоминируемые решения – Парето-оптимальный фронт – многокритериальной оптимизационной задачи взвешивания и отбора признаков представлены на рис. 4.

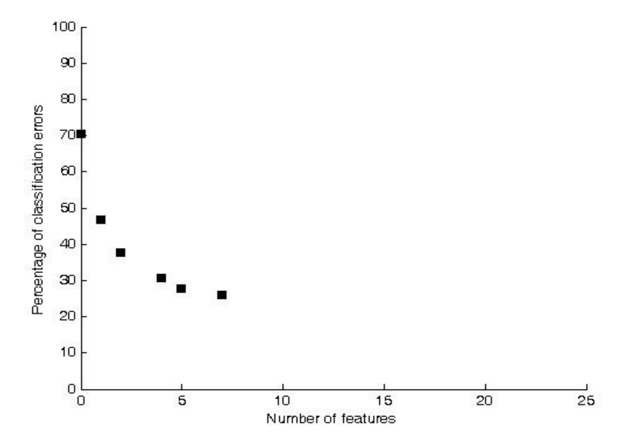


Рисунок 4 – Недоминируемые подмножества признаков: горизонтальная ось определяет количество признаков, вертикальная – ошибку классификации (распознавания) подтипов ТИА

Весовые коэффициенты недоминируемого подмножества, обеспечивающего минимальную ошибку классификации 25,7 %, представлены в табл. 2.

Таблица 2 – Весовые коэффициенты отобранных признаков

Признак	PROFESSN	HERED_CV	HYPERTEN	CORCARSC
Вес	3,7	3,3	5,4	6,5
Признак	BRONCHRO	HEADACHE	VERTIGO	
Вес	9,6	4,2	9,4	

Заключение

В настоящей работе описаны два способа выделения признаков, которые позволяют сократить сложность и повысить точность классификации путем получения недоминируемых подмножеств признаков с использованием генетического алгоритма, предназначенного для решения многокритериальных задач. Преимуществом использования ГА в этом случае является получение нескольких решений с последующей возможностью привлечения знаний и опыта экспертов с целью выбора окончательного подмножества предсказывающих признаков. Применение предложенного эволюционного подхода к дифференциальной диагностике подтипов транзиторных ишемических атак позволяет сконцентрировать внимание на небольшом количестве признаков, являющихся в этом случае наиболее информативными, и получить классификационное решение, которое не уступает по точности классификации с решением, полученным с учетом 25 исходных признаков.

Дальнейшим направлением исследований является использование ГА для параллельного отбора признаков и наблюдений из набора данных, что позволит сократить количество прототипов, используемых при проведении классификации методом *k*-ближайших соседей. Для больших наборов данных такой отбор позволит уменьшить временные и вычислительные затраты на осуществление классификации объектов

данных. Интерес представляет также использование результатов применения эволюционного подхода к отбору признаков для построения ансамблей классификаторов и применение различных комбинационных методов для получения более высокой точности классификации [8].

Литература

1. Dash M. Feature selection for classification // Intelligent Data Analysis. – 1997. – Vol. 1, № 3. – P. 131-156.
2. Дривотинов Б.В. Адаптивная нейро-нечеткая модель для дифференциальной диагностики подтипов транзиторных ишемических атак // Военная медицина. – № 4. – 2007. – С. 101-106.
3. Deb K. Multi-Objective Optimization using Evolutionary Algorithms // John Wiley & Sons, England. – 2001.
4. Cover T.M. Nearest neighbor pattern classification // IEEE Transactions on Information Theory. – 1967. – Vol. 13, № 1. – P. 21-27.
5. Goldberg D. Genetic algorithms in search, optimization and machine learning – Reading (MA): Addison-Wesley, 1989. – 432 p.
6. Siedlecki W. A note on genetic algorithms for large scale feature selection // Pattern Recognition Letters. – 1989. – Vol. 10, № 5. – P. 335-347.
7. Dimensionality reduction using genetic algorithms // IEEE Transactions on Evolutionary Computation. – 2000. – Vol. 4, № 2. – P. 164-171.
8. Kittler J. & Roli F. Multiple Classifier Systems // Proc. of 2nd International Workshop, MCS2001, Cambridge (UK) 2-4 July 2001, Lecture Notes in Computer Science. – Vol. 2096. – Springer-Verlag, Berlin.

Н.А. Новосьолова, О.С. Мастикін, І.Е. Том

Еволюційний підхід до виділення інформативних ознак у завданнях аналізу медичних даних
У статті розглядається підхід до виділення інформативних ознак стосовно до завдання розпізнавання підтипів транзиторних ішемічних атак. Згідно із запропонованим підходом завдання виділення ознак розглядається як завдання багатокритеріальної оптимізації із двома критеріями. Оптимізація здійснюється з використанням спеціального генетичного алгоритму, що дозволяє в процесі еволюції одержати безліч недомінуючих рішень оптимізаційної задачі. Запропонований підхід дозволяє підключити експерта на етапі остаточного прийняття рішень, надаючи йому можливість відбору підмножини ознак, найбільш відповідного його знанням і уявленням про розв'язуване завдання.

N.A. Novoselova, A.S. Mastikin, I.E. Tom

Evolutionary approach to informative feature extraction in medical data analysis
The paper proposes an approach to informative feature extraction as applied to recognition of transient ischemia attack subtypes. According to the approach the feature extraction is considered as multi-objective optimization task with two criteria. The optimization process is performed with special genetic algorithm, allowing to find the set of non-dominated solutions of optimization task during evolution. The proposed approach enables the attraction of medical expert to final decision making, taking into account his knowledge and clear idea of medical task.

Статья поступила в редакцию 29.07.2008.